Published in final edited form as:
*Vis cogn*. 2009 August 1; 17(6-7): 1185–1204. doi:10.1080/13506280902978477.

# Modelling the role of task in the control of gaze

**Dana H. Ballard** and **Mary M. Hayhoe**
Center for Perceptual Systems, University of Texas, Austin, TX, USA

## Abstract

Gaze changes and the resultant fixations that orchestrate the sequential acquisition of information from the visual environment are the central feature of primate vision. How are we to understand their function? For the most part, theories of fixation targets have been image based: The hypothesis being that the eye is drawn to places in the scene that contain discontinuities in image features such as motion, colour, and texture. But are these features the cause of the fixations, or merely the result of fixations that have been planned to serve some visual function? This paper examines the issue and reviews evidence from various image-based and task-based sources. Our conclusion is that the evidence is overwhelmingly in favour of fixation control being essentially task based.

### Keywords

Gaze control; Saliency; Task modeling; Saccades

Yarbus's original work in understanding gaze recordings (Yarbus, 1967) in the 1950s and 1960s revealed the enormous importance of gaze in revealing the underlying structure of human cognition. In his most compelling demonstration, he showed that a subject viewing a painting responded with markedly different eye fixation patterns when asked different questions about the image. Although earlier work by Buswell and Dodge (Buswell, 1935; Erdmann & Dodge, 1898) had implied such cognitive influences on the choice fixations, Yarbus's demonstrations left no doubt about the role of cognition directing gaze.

From this perspective of this pioneering work, it is somewhat surprising that the first significant computational theory of vision (Marr, 1982) avoided the study of gaze as well as any influence of cognition on the extraction of information from the retinal array. In his "principle of least commitment", Marr argued the case for the role of the cortex in building elaborate goal-independent descriptions of the physical world. Marr was no doubt influenced by the groundbreaking work of Hubel and Weisel (1962), who showed that striate cortex was organized into a retinotopic map of the visual world, centred on the point of gaze. Subsequent work revealed that all of visual cortex was hierarchically organized into a series of retinotopic maps containing ever more abstract properties of the visual world. At almost the same time work in visual search revealed a stunning difference in search times between displays containing several items with just one feature defining differences between items and displays with conjunctions of two features defining the difference (Treisman & Gelade, 1980), suggesting that groups of features at retinotopic locations was a natural way of organizing the visual stimulus.

Please address all correspondence to Dana H. Ballard, Center for Perceptual Systems, SEA 4.328A, 1 University Station, A8000, The University of Texas, Austin, TX 78712-0187, USA. dana@cs.rochester.edu.

As a consequence of the focus on retinotopy, when researchers took on the task of defining computational mechanisms for directing gaze deployment, these turned out to be predominantly image based. Koch and Ullman defined the saliency map: A retinotopic accounting of different retinotopic organizations of specific image features such as colour, texture, and motion (Itti & Koch, 2001; Koch & Ullman, 1985). Retinotopic locations rich in such features were calculated to be salient and potential fixation points. Subsequent additions allowed these locations to be modulated to account to different tasks (Navalpakkam & Itti, 2003; Wolfe, 1994) and the statistics of such features (Itti & Baldi, 2005), and the concept of the saliency map has become a central organizing focus of models of gaze control.

Saliency theories have been compelling, but have many drawbacks. They usually cannot predict the exact fixation points and can leave more than half of the fixations unaccounted for (e.g., Foulsham & Underwood, 2008). It seems likely that the central problem is that they are correlated with fixation behaviours but not their cause. This point has been made by Einhauser, Rutishauser, and Koch (2008), Henderson (2007), and Tatler (2007). Other compelling reasons for this are (1) the dominating role of cognitive goals dictates many image calculations that cannot be expressed in terms of the saliency of conjunctions of features and (2) the situated nature of the human visual system's limited view in a three-dimensional world means that many fixation targets are remembered locations that are not visible when the saccade is initiated.

Recent work has begun to tackle the problem of describing a theory that accounts for the role of the cognitive process and spatial environs that are controlling the subject's behaviours, and there is mounting evidence that such a theory must be central. Experiments using lightweight head-mounted eyetrackers show that fixations are extracting very specific information needed by the human subject's ongoing task (Droll, Hayhoe, Triesch, & Sullivan, 2005; Jovancevic, Sullivan, & Hayhoe, 2006; Triesch, Ballard, Hayhoe, & Sullivan, 2003). The task context introduces enormous economies into this process: If a subject needs to locate a red object near a blue object, the search for that object can be limited to just blue portions of the image followed by a local search for red; vast amounts of extraneous detail can be neglected (Ballard, Hayhoe, Pook, & Rao, 1997; Roelfsema, Khayat, & Spekreijse, 2003; Swain & Ballard, 1991). The visual information-gathering component of almost every task will introduce similar economies.

The purposes of this paper are twofold. We first selectively review research on the deployment of gaze with the view to highlighting difficulties in saliency models. Next we introduce a cognitive model for directing fixations based on learned reward. Rather than offer a *rapprochement* between image-based and task-based approaches, we continue an earlier argument advanced by Ullman (1984) in his classic visual routines paper. That is, that the cognitive processes that operate on image data in order to support cognition have a fundamentally different character than the image structures they use in that process.

## EVIDENCE FROM SPORTS

One obvious venue for studying eye movements is sports. Given the time-critical coordination of movements involved, one would suspect that this is a case where eye movements are entirely task based and in fact this turns out to be the case. In classic studies of cricket, Land has shown that batters fixate the area on the pitch where the ball is expected to bounce after first fixating the pitch release to get information on the ball's upcoming trajectory. The best batter's gaze change tends to be about 100 ms before those of average batters (Land & McLeod, 2000). This venue provides an ideal setting to contrast the information gained from saliency methods with information gained by understanding the

task. The bowler is of course very salient so it would be easy to produce a high valued saliency measure for the visual area containing him, although perhaps more difficult to highlight the crucially important hand area. On the other hand the area where the ball bounces is virtually featureless, gaining its importance from the surrounding context and the batter's prior knowledge of where pitches typically land. Thus, a saliency model would be able to predict half the fixations but would miss the essence of them. The first fixation's purpose is to measure information that would predict the landing point and spin of the ball; the second fixation is to plan the batter's response.

The use of fixation to plan motor responses in saliency-poor areas is ubiquitous in ball sports. Land and McLeod (2000) have shown that players fixate the predicted bounce point in table tennis, and Hayhoe (McKinney, Chajka, & Hayhoe, 2008) has shown that players fixate the featureless front wall for squash returns. Back wall returns are even more impressive; players fixate a point in empty three-dimensional space that is predicted to be the contact point between the ball and the racquet. In another venue where normal, unskilled subjects have to catch a bounced ball, subjects fixate a three-dimensional point above the bounce point. None of these points could be predicted by a saliency model as the local image features are meaningless for these tasks. Of course the path of the ball, the placement of the player/catcher and their positions in three-dimensional space are all important, but none of this information is part of saliency models.

## EVIDENCE FROM COMPLEX TASKS

Sports are interesting in part because the performance of the players is time critical, so they leave open the possibility that in other normal behaviours that are not so demanding, the use of fixations could be less predictable with task information. However, all the emerging evidence implies that this is not the case. It is far more likely that each fixation has a specific purpose even when the observer may not be conscious of it. In studies of car following in a virtual environment, Shinoda, Hayhoe, and Shrivastava (2001) showed that the percentage of fixations on a lead car dropped from 75% to 43% when subjects had to also pay attention to intersection signs. Although it might be possible to adjust a task-based modulation of a saliency map to account for these changes the difficulties are formidable. The visual targets vary hugely in scale as the driver progresses and subjects tend to look at signs at different times.

In a classic experiment, subjects copied patterns of coloured blocks on a display, moving blocks used in the copy from a reservoir of extra blocks with a computer cursor. In this task the modal behaviour was to fixate the pattern to choose a block, remember its colour, find and select the block to be used in the copy, and then use another fixation of the pattern to determine the placement of the selected block in the copy. Even though subjects could obtain both pieces of information (position, colour) with a single fixation, they preferred to use two separate fixations, presumably with the goal of avoiding the carrying cost of remembering the relative position for as long as possible. Thus the first fixation was to determine the colour of the block in turn to compute the subsequent fixation to the reservoir, and the second fixation was to obtain the relative position of the block in the pattern (Hayhoe, Bensinger, & Ballard, 1998).

The focused use of fixations in the block copying task is also a hallmark of subsequent experiments studying natural behaviour in that knowing the task allows the purpose of the fixation to be understood. We will describe these experiments to illustrate this point but before we do let us return to the issue of saliency maps to illustrate their conundrums in this case. One could certainly produce a saliency map for the blocks images and compute points of saliency, but without understanding the task at hand it would be virtually impossible to

predict the sequence of eye movements for this task as the static images contain no information as to which salient location is to be preferred over any other.

Another complex task studied by Hayhoe is that of making a peanut butter and jelly sandwich and filling a cup with Coke (Hayhoe, Shrivastrava, Myruczek, & Pelz, 2003). This is a common everyday task but is laden with complexities in eye–hand coordination that challenge the notion of saliency. For example in placing peanut butter on the bread, subjects take advantage of the fact that peanut butter reliably sticks to the knife and make a gaze targeting fixation to the point on the bread where the tip of the knife is to end up to begin spreading. In contrast, jelly is less viscous and more precarious on the knife and thus is guided to the bread with a pursuit eye movement. This kind of knowledge, which is ingrained in any sandwich maker, is way beyond the reach of the capabilities of saliency maps as the knowledge is simply not image based.

At this point it might be germane to discuss one of the ways proposed to extend saliency maps and that is to modulate them with task information. If the information in tasks can be related to image features then, the features themselves can provide a basis for modulating the saliency map. If it is known that jelly is an important component of the task, then one can increase the weight of jelly's dark purple colour in the saliency computation, thus biasing the choice of fixation points. There are lots of difficulties making this work in general but the one that is appropriate for discussion here is that jelly is only important at a few points in the overall task: Finding the jelly jar, extracting the jelly from the jar, and replacing the jelly jar's lid. Thus, any method for routinely increasing saliency would create false targets for all the other moments of sandwich construction and drink pouring. Of course one could introduce knowledge of possible sequences of sandwich making and selectively enhance purple when jelly was important but at this point the idea of task knowledge has been ceded.

Another point that proves problematic for saliency is that of task-based memory. We illustrate this point with another copying example, this time from Aivar, Hayhoe, and Mruczek (2005). Subjects copy a toy model in a virtual environment. The toy is made from German Baufix parts that subjects manipulate with a cursor, as in the previously described block copying task. The important differences here are that (1) subjects copy the same toy repeatedly so that they can learn where the parts are in the reservoir and (2) the distance between the construction site and reservoir is such that when adding parts to the construction, the reservoir is not in view. Nonetheless when subjects have repeated the task they are able to make saccades to the point in virtual three-dimensional space at which the next part is suspended. This is tested by moving the placement of the part when not in view and observing that the initial saccade goes to the vacated spot. Of course this cannot be handled by saliency. In the first place the image is not available at the beginning of the saccade, but even if it were the part is no longer at that position so the resultant saliency of that location is zero. One could try to overcome this difficulty with baroque manipulations of the saliency map, but they would only be shoehorning implicit knowledge of the task into the image-based representation. Furthermore there are other issues not included at this point which are about to be addressed.

Since many aspect of ordinary visually guided behaviour are clearly dominated by the information that is required for the momentary visual operation, what is the potential role of examining the properties of the stimulus as a basis for gaze behaviour? One of implicit rationales might be that tasks are special in some way, and that there is some body of visual processing that does not involve tasks. Consequently, many of the experiments involve what is called "free viewing", with the goal of isolating task-free visual processing. It is possible that the global visual perception of a scene is distinct in some way from the kind of vision

involved in tasks. Certainly humans need to extract information about the spatial structure of scenes and the identity of the objects, but is this qualitatively different from specific visual operations such as those involved in visual search, or extracting location information for grasping an object, operations that are performed in the context of a task such as making a sandwich? The suggestion in the next section is that all vision can be conceptualized as a task of some kind. The issue is important and needs to be examined explicitly. What we think of as "seeing" is the consequence of extensive experience in the visual environment during development (Geisler, 2008) and the extraction of information such as gist presumably reflects not only passive visual experience but also the constraint that this information is useful for the organism in some way. Thus it is hard to logically separate the effects of stimulus from the effects of task.

Another important assumption that needs to be examined is that fixation patterns in two-dimensional photographic renderings of a scene will be the same if the observer were actually in that scene. Although this might be the case in some instances, there is no guarantee that it will be true. Real scenes are three-dimensional, and the image changes as a consequence of inevitable body movements. The scale of a rendered image is typically different from the scale of the image if the subject were actually in that scene.

## EVIDENCE FOR VISUAL ROUTINES

The discussion up to this point has focused on what might be termed macroscopic issues with respect to eye fixations, that is, specifying their targets broadly and issues as to the overall task context in influencing those choices. But a host of other issues emerge with respect to more microscopic issues, namely the detailed computations that are done during fixation. A variety of experiments have indicated that the visual information acquired during a fixation may be quite specific. In an experiment by Ballard, Hayhoe, and Pelz (1995), observers copied simple coloured block patterns on a computer screen, by picking up blocks with the mouse and moving them to make a copy. In the course of copying a single block, subjects commonly fixated individual blocks in the model patterns twice, once before picking up a matching block, and once before placement. Given the requirements of the task, a reasonable hypothesis is that block colour is acquired during the first fixation, and the next fixation on the block is to acquire its location. A subsequent experiment where changes were made to the block colours at different stages of the task supported the interpretation that the first and second fixations on a model block subserved different visual functions (Hayhoe et al., 1998; Hoffman, Landaub, & Pagani, 2003). Further evidence that fixations are for the purpose of extracting quite specific information is given by Droll et al. (2005), who found that subjects are selectively sensitive to changes made in task relevant features of an object they were manipulating, even though they fixated the object directly for several hundreds of msec. Triesch et al. (2003) also found selective sensitivity to task relevant changes in a manipulated object. This suggests that many simple visual computations involve the ongoing execution of special-purpose "visual routines" that depend on the immediate behavioural context, and extract only the particular information required at the moment. The idea of visual routines was first introduced by Ullman (1984). The essential property of a routine is that it instantiates a procedure for acquiring specific information called for by the current cognitive agenda. Selection of just the task specific information from a scene is an efficient strategy. Task specific strategies not only circumscribe the information that needs to be acquired, but also allow the visual system to take advantage of the known context to simplify the computation (Ballard et al., 1997). This selective acquisition may be reflected in even low-level cortical areas whose neural activity depends not only on stimulus features but on task context (Ito & Gilbert, 1999; Roelfsema, Lamme, & Spekreijse, 1998).

To illustrate the concept of a visual routine, we consider the task of filling a cup with coke. While pouring the coke subjects lock their gaze on the level of coke and track its progress towards the rim. Each subject has a preferred level that he or she can duplicate repeatedly. The obvious conjecture is that subjects are using a template matching approach whereby they are mentally matching a "filled coke cup" template against the current image, stopping when a match criterion is achieved. We have shown with a model in virtual environment that this simple information is adequate for performing the task, and can reproduce the standard deviation of fill levels of a given subject using template matching. Thus, the suggestion is that vision is composed of specialized computations of this kind.

Consider the problem the cup-filling example poses for saliency. Much of the context for the visual routine is provided by the body itself. Since the subject's hands are holding the cup and coke bottle, proprioception can provide the essential geometric information for filling the cup. The weight of the filling cup is another cue. Vision is just needed to detect the final condition of a filled cup. In this venue there are two problems. The first is that the contrast between the liquid and cup colour can be regulated so that their impact on the saliency computation can be reduced to near zero. Thus, without extensive priming the level of the fluid in the cup will be invisible. Second, the subject's gaze tracks the filling level for the duration of the filling process (it is likely that the motion of the fluid relative to the cup is used in doing this, based on a patient studied by Zihl, von Cramon, & Mai, 1983, who had a specific motion deficit and had trouble filling cups). If the knowledge of the task is provided then this behaviour is reasonable, but absent it, there is no reason for using gaze in this way and no way to predict the behaviour purely on the basis of saliency.

A challenging example for saliency models in neuroscience comes from Roelfsema et al.'s (1998) primate studies. In his experimental setup a monkey has to fixate a central point and then on command, make a saccade to one of two radial lines projecting outward from the fixation point. The line that must be chosen is the one that is attached to the fixation point. The experiment takes advantage of the fact that in programming a saccade from a cue onset takes on the order of 250–300 ms so during that time one can record from fixed retinotopic locations in cortical areas such as striate cortex (V1). The experiments show that simple cells on along the line's path increase their firing pattern at a time commensurate with the hypothesis that the monkey solves the task of defining the saccade target by mentally tracing the length of the line to the required end point. The task was made harder by replacing the attachment condition with colouring the fixation point and stubs at the near ends of the lines. The line to be traced is the one that now has the same colour stub as the fixation point. The elevated simple cell response now occurs later in time, consistent with the hypothesis that the monkey now solves two tasks. The main point here for our focus is that line tracing is a technically difficult problem that is outside the domain of static saliency models.

A final example of the use of visual routines posing difficulties for saliency comes from Droll and Hayhoe (2007). In a virtual block task, subjects look a block they are manipulating at different times, but some of those fixations are to obtain its features (e.g., colour), whereas others are to follow task instructions. The point is that the fixation is on the block in both cases but the actual detailed processing that the visual system is doing is very different. The saliency map cannot distinguish these two without having a detailed task model.

## HUMAN VISUAL SEARCH

An important aspect of saccadic eye movements that has implications for saliency is their use in visual search. Understanding this venue draws upon the development of reverse correlation in the computation of the search target used by subjects. Given a succession of searches for a target in noise, the experimenter can keep a record of the subject's false

positives and true positives and average these in order to produce an image template for the searched target. For example, it has been demonstrated that for searching for targets under low signal-to-noise conditions that the features extracted are often idiosyncratic and not easily related to saliency axes (Rajashekar, Bovik, & Cormack, 2007).

In a related approach, Geisler (2008) asks the question of what is an optimal search pattern for a target embedded in noise given that the retinotopic resolution heavily emphasizes the fovea; a model of the search process predicts subjects' performance accurately. These results have been replicated experimentally by Caspi, Beutter, and Eckstein (2004).

Rao, Zelinsky, Hayhoe, and Ballard (2002) studied a condition where subjects had to search a natural scene such as a tabletop image where one to five objects might appear. Just prior they were shown an image of the target object. There were two conditions, one where at the outset the subjects were given a short view of the tabletop and could memorize the locations of the objects and another where the preview was absent. The instructed response was target present or absent indicated by an appropriate keypress. Although eye movements were not controlled beyond an initial fixation, subjects invariably fixated the target in the course of the response. However, as shown in Figure 1, in the preview condition, subjects usually fixated the target with one saccade, whereas in the no preview condition the modal number of saccades was three.

All these approaches again raise problems for the saliency model. Since the search target is specified by the experimenter, it cannot be a ready product of the priors that saliency computations assume. Of course the actual correlation-based computations that are done can be embraced as saliency, but that would defeat the fundamental stance of saliency as a method of filtering the image ab initio to delimit possible fixation targets. A very nice paper nonetheless blurs this distinction; Navalpakkam and Itti (2007) show that when the search task is cast in terms of artificial distinctions of primitive features, the task can be described in terms of modulations of such features and, furthermore, human subjects obey the dictates of signal detection theory. Interestingly, the paper does not exhibit eye movements, perhaps because the task is done in a limited part of the visual field.

## MODELLING TASK-DIRECTED FIXATIONS

We have made the case that the main source of explanations of fixation locations is not image saliency but rather latent cognitive variables. In this case the task becomes describing the complex human cognitive system in a way that its descriptive components can be related to fixations. This is not an easy task to do without making substantial claims about the workings of cognition interacting with the visual system.

One such system that tackles the cognition–action interface is that of Sprague, Ballard, and Robinson (2007). The central assumption they make is that the system can be composed of modularized sensorimotor behaviours which can cooperate in small sets without interference from each other. For realizing compositions of modular behaviours, following work in psychology and robotics (e.g., Bonasso, Firby, Kortenkamp, Miller, & Slack, 1997), they develop an abstract cognitive architecture composed of three levels: *Central executive*, *arbitration*, and *behaviour*. The central executive level of the hierarchy maintains an appropriate set of active behaviours from a much larger library of possible behaviours, given the agent's current goals and environmental conditions. The composition of this set was evaluated at every simulation interval, taken to be 300 ms. The arbitration level addresses the issue of managing competing active behaviours. Thus, an intermediate task is that of mapping action recommendations onto the body's resources. Since the active behaviours must share perceptual and motor resources, there must be some mechanism to arbitrate their

needs when they make conflicting demands. The behaviour level describes distinct jobs that are necessary, such as interrogating the image array in order to compute the current state.

The models for each of these levels are implemented and tested on a human avatar. The virtual human vision avatar has physical extent and programmable kinematic degrees of freedom that closely mimic those of real humans as well as software for modelling the physics of collisions. This software base has been augmented with our control architecture for managing behaviours. Each behaviour has a very specific goal, and contains all the structure for the extraction of information from visual input that is in turn mapped onto a library of motor commands. Figure 2A shows the avatar in the act of negotiating a pavement that is strewn with obstacles (blue objects) and litter (purple objects) on the way to crossing a street. Figure 2B shows a human subject in the same environment.

A central problem for task-directed vision concerns the deployment of gaze. The small fovea makes its use to obtain accurate measurements a premium. So in the case of multiple active behaviours, which of them should get the gaze vector at any instant? An elegant solution to this problem is to calculate the amount each behaviour stands to gain by updating its state. Where $Q(s_i,a)$ is the discounted value of behaviour $i$ choosing an action $a$ in state $s_i$, an agent that chooses an action that is suboptimal for the true state of the environment can expect to lose some reward, estimated as follows:

$$\text{loss}=E\left[\max_a \sum Q_i(s_i,a)\right]-E\left[\sum Q_i(s_i,a_E)\right] \quad (1)$$

The term on the left-hand side of the minus sign expresses the expected return if the agent were able to act with knowledge of the true state of the environment. The term on the right expresses the expected return if forced to choose an action based on the state estimate. The difference between the two can be thought of as the cost of the agent's current uncertainty. The total expected loss does not help to select which of the behaviours should be given access to perception. To make this selection, the loss value can be broken down into the losses associated with the uncertainty for each particular behaviour $b$:

$$\text{loss}_b=E\left[\max_a\left(Q_b(S_b,a)+\sum_{i\in B,i\neq b} Q_i^E(S_i,a)\right)\right]-\sum_i Q_i^E(S_i,a_E) \quad (2)$$

The expectation on the left is computed only over $s_b$. This value is the expected return if $s_b$ were known, but the other state variables were not. The value on the right is the expected return if none of the state variables are known. The difference is interpreted as the cost of the uncertainty associated with $s_b$. This calculation is for all the active behaviours and the one that has the most to lose gets the vector. Figure 3 shows this happening for a walking segment.

As Figure 3C shows, the improvement is small, but nonetheless highly significant. The narrow margin highlights the difficulty of reward-based hypotheses about eye fixations that would attempt to have the results of the fixation directly alter the Q-table. The relative value of any given fixation is small enough so as to be practically undetectable by the learning process. The simulations were unable to detect any systematicity in these variations, and led us to propose the most-to-gain model which uses the state table to calculate the value of an eye fixation, but does not attempt to adjust the Q-table otherwise.

The simulations can be thought of as tackling the problem of *when* gaze is deployed, but we are also interested in *where* gaze is deployed. Figure 4 shows a case where we obtained Laurent Itti's saliency program and used it to calculate salient points in the walkway setting. Since subjects had walked in this setting their subjects' fixations could be compared directly with those predicted by the program. The program does not return a single point, but a spatial distribution of possible fixation points so some criterion has to be used for selecting a fixation. If the saliency distribution overlapped the object selected by a subject, the result was labelled a "match"' otherwise the label was "no match". The figure shows representative results. Even in this simple setting less than half of the fixations can be accounted for with saliency. In contrast, the model uses task-directed visual routines based on human performance (Rothkopf, Ballard, & Hayhoe, 2007), so that the landing sites of the routines are qualitatively accurate: Litter fixations are to the centre of the litter; obstacle fixations land on the furthest edge; pavement fixations land on the pavement edge. The reader should compare Figure 3B with Figure 4B and C.

## MODELLING TASKS WITH SEQUENTIAL STEPS

In the walking example each behaviour has very a minimal state description. For example, staying on the pavement just requires measuring the pavement edge. The history of the traverse is not needed. However, more complicated behaviours require much more elaborate internal state descriptions. Specifying the details of those descriptions is challenging ongoing research enterprise and is taking many directions. Herein we briefly describe our own work but one could just as easily use other examples such as Nytrøm and Holmqvist (2008) and Oliva and Torralba (2006) as illustrations. The point is that the surface image manipulations are just the tip of an the iceberg of representational structure needed to interpret fixation choices. Furthermore, our example only addresses the recognition issues in interpreting observed fixations. Additional structure is needed to generate the fixations in the process of producing the behaviour.

Consider the process of making a peanut butter sandwich (Hayhoe et al., 2003). If you want to put peanut butter on a slice of bread, you must be holding the knife and you must have taken the lid off the peanut butter jar. Modelling this state is not straightforward owing to a number of factors. Consider the problem of watching someone make a sandwich and describing what has transpired. The basic actions must be measured and recognized. However, all the steps in the process are noisy and hence the description must necessarily be probabilistic. Now consider describing the order of steps making a sandwich. Since there are over 1000 distinct ways of making it that differ in the order of the steps, any particular sequence of steps is best described probabilistically. A central way of handling probabilistic information goes under the name *graphical models*. These are particularly valuable when the basic dependencies are in the form of conditional probabilities, as in the sandwich-making case. Although some care has to be taken in developing a graphical model, Yi and Ballard (2006) were able to do it. This is a very demanding task, since the model must take head, hand, and eye data from the subjects and, at any given time, recognize what stage in the sandwich making is occurring, as shown in Figure 5.

For this task the graphical model is in the form of a Bayes Net. Such a network is a suitable tool for this class of problems because it uses easily observable evidence to update or infer the probabilistic distribution of the underlying random variables. A Bayesian net represents the causalities with a directed acyclic graph, its nodes denoting variables, and edges denoting causal relations. Since the state of the agent is dynamically changing and the observations are being updated throughout the task execution process, one needs to specify the temporal evolution of the network. Figure 6 illustrates the two slice representation of a Dynamic Bayes Network (DBN). Shaded nodes are observed; the others are hidden.

Causalities, represented by straight arrows, are determined by probability distribution matrices.

Each of the states can take on several discrete values as shown by the Tables 1 and 2. Visual and motor routines produce specific values for each of the shaded nodes and the standard Bayes Net propagation rules fill in values for the task nodes. The state of the lowest hidden node is determined by its prior distribution in the first time/slice and thereafter jointly determined by its previous state and the transition matrix, as denoted by the curved arrow shown in Figure 6.

The two-slice representation can be easily unrolled to address behaviours with arbitrary numbers of slices. At each moment, the observed sensory data (grey nodes), along with its history, are used to compute the probability of the hidden nodes being in certain states:

$$p(S^t|O^{[1,t]}) = P(S^1)P(O^1|S^1)\prod_{t=2}^{t} P(S^t|S^{[t-1]})P(O^t|S^{[t]})$$

where $S^t$ is the set of states of hidden nodes at time $t$, $O^{(1,t)}$ is the observations over time span $(1,t)$. Behaviour recognition computes the states of each hidden node $S^t$ at time $t$ that maximize the probability of observing the given sensory data:

$$S^t = \arg\max_S P(S^t) = S|O^{[1,t]})$$

The point of this elaborate example is simply that all the key variables that direct the progress of interpreting sandwich making are part of an estimate of the sandwich constructor's cognitive program. Image data is important, along with hand measurements, but primarily for conforming hypotheses in the cognitive program. The image structure is not the *cause* of the sandwich being made.

## CONCLUSION

The first goal of this paper was to show that it is very unlikely that the saliency map could be the cause of gaze changes. The principal evidence is that almost all behaviour is goal oriented and the object of these goals does readily translate into constellations of image features in a significant number of cases. Thus, these cases are not capable of being modelled as saliency map targets. This is not to say that the saliency map is not without value as in many other cases the features in the saliency map can be used to compute the planned point of fixation. However, the main point still remains that the cause of such a computation comes from the latent variables associated with the subject's internal goals and not directly from the image itself.

In racquet sports where the object is to hit a ball, the position that the racquet must meet the ball is most often a proximal point in three-dimensional space that is determined by distal information. In completing a complex task, often the target of a fixation depends on remembered information obtained on prior fixations and not on the current image. In complex tasks the fixation point can depend on a computation that depends on the image features that cannot be anticipated without knowledge of the task itself.

The conclusion of all these observations is that in order to progress, a substantial effort must be invested in modelling tasks so to have the variables used in computing fixation points made explicit. As an example, we described a way of composing behaviours whose

components are learnt using reinforcement. Such behaviours can be used to generate fixations on the basis of rewarding reduction in uncertainty. This idea has at the moment the status of a conjecture, but nonetheless illustrates the motivation for a nonsaliency theory of gaze control.

Finally, to illustrate the possibility of using the information acquired at the point of gaze with that of other body actions to guide behaviour, we showed that the steps in a complex behaviour such as sandwich making could be recognized with just that information as input. Again the status of these variables is provisional, but nonetheless they constitute an existence proof that this sparse information is sufficient to accomplish the task. A huge amount of additional work will need to be done before one could safely establish the role of cognition in gaze control, but the aim of this paper is to argue that this research is necessary and will supplant strictly image-based computational models.

## Acknowledgments

This work was supported by NIH grants RR02983, MH60624, and EY05729. The authors gratefully acknowledge the assistance of the reviewers whose detailed comments and suggestions greatly improved the paper.

## References

Aivar P, Hayhoe M, Mruczek R. Role of spatial memory in saccadic targeting in natural tasks. Journal of Vision 2005;5:177–193. [PubMed: 15929644]

Ballard D, Hayhoe M, Pelz J. Memory representations in natural tasks. Journal of Cognitive Neuroscience 1995;7:66–80.

Ballard D, Hayhoe M, Pook P, Rao R. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 1997;20:723–767. [PubMed: 10097009]

Bonasso R, Firby R, Kortenkamp D, Miller D, Slack M. Experiences with an architecture for intelligent reactive agents. Journal of Experimental and Theoretical Artificial Intelligence 1997;9:237–256.

Buswell, GT. How people look at pictures. Chicago: University of Chicago Press; 1935.

Caspi A, Beutter B, Eckstein M. The time course of visual information accrual guiding eye movement decisions. Proceedings of the National Academy of Sciences of the USA 2004;101:13086–13090. [PubMed: 15326284]

Droll J, Hayhoe M. Deciding when to remember and when to forget: Trade-offs between working memory and gaze. Journal of Experimental Psychology: Human Perception and Performance 2007;33(6):1352–1365. [PubMed: 18085948]

Droll J, Hayhoe M, Triesch J, Sullivan B. Task demands control acquisition and maintenance of visual information. Journal of Experimental Psychology: Human Perception and Performance 2005;31(6): 1416–1438. [PubMed: 16366799]

Einhauser W, Rutishauser U, Koch C. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. Journal of Vision 2008;8:1–19.

Erdmann, B.; Dodge, R. Psychologische Untersuchungen uber das Lesen. Halle, Germany: N. Niemeyer; 1898.

Foulsham T, Underwood G. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision 2008;8:1–17. [PubMed: 18318632]

Geisler WS. Visual perception and the statistical properties of natural scenes. Annual Review of Psychology 2008;59:167–192.

Hayhoe M, Bensinger D, Ballard D. Task constraints in visual working memory. Vision Research 1998;38:125–137. [PubMed: 9474383]

Hayhoe M, Shrivastrava A, Myruczek R, Pelz J. Visual memory and motor planning in a natural task. Journal of Vision 2003;3:49–63. [PubMed: 12678625]

Henderson J. Regarding scenes. Current Directions in Psychological Science 2007;16:219–227.

Hoffman J, Landaub B, Pagani B. Spatial breakdown in spatial construction: Evidence from eye fixations in children with Williams syndrome. Cognitive Psychology 2003;46:260–301. [PubMed: 12694695]

Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology 1962;160(1):106–154. [PubMed: 14449617]

Ito M, Gilbert G. Attention modulates contextual influences in the primary visual cortex of alert monkeys. Neuron 1999;22:593–604. [PubMed: 10197538]

Itti L, Baldi P. Bayesian surprise attracts human attention. Proceedings of Neural Information Processing Systems 2005;19:547–554.

Itti L, Koch C. Computational modeling of visual attention. Nature Reviews Neuroscience 2001;2:194–203.

Jovanevic J, Sullivan B, Hayhoe M. Control of attention and gaze in complex environments. Journal of Vision 2006;6:1431–1450. [PubMed: 17209746]

Koch C, Ullman U. Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology 1985;4:219–227. [PubMed: 3836989]

Land MF, McLeod P. From eye movements to actions: How batsmen hit the ball. Nature Neuroscience 2000;3:1340–1345.

Marr, D. Vision. New York: Henry Holt & Co; 1982.

McKinney T, Chajka K, Hayhoe M. Pro-active gaze control in squash [Abstract]. Journal of Vision 2008;8(6):111a.

Navalpakkam V, Itti L. Modeling the influence of task on attention. Vision Research 2005;45:205–231. [PubMed: 15581921]

Nytrøm M, Holmquist K. Semantic override of low-level features in image viewing—both initially and overall. Journal of Eye Movement Research 2008;2:1–11.

Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research 2006;155:23–36. [PubMed: 17027377]

Rajashekar U, Bovik A, Cormack L. Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. Journal of Vision 2007;6(4):379–386. [PubMed: 16889476]

Rao R, Zelinsky G, Hayhoe M, Ballard D. Eye movements in iconic visual search. Vision Research 2002;42:1447–1463. [PubMed: 12044751]

Roelfsema P, Khayat PS, Spekreijse H. Subtask sequencing in the primary visual cortex. Proceedings of the National Academy of Sciences USA 2003;100:5467–5472.

Roelfsema P, Lamme V, Spekreijse H. Object-based attention in the primary visual cortex of the macaque monkey. Nature 1998;395:376–381. [PubMed: 9759726]

Rothkopf C, Ballard D, Hayhoe M. Task and scene context determines where you look. Journal of Vision 2007;7(14):1–20. [PubMed: 18217811]

Shinoda H, Hayhoe M, Shrivastava A. Attention in natural environments. Vision Research 2001;41:3535–3546. [PubMed: 11718793]

Sprague N, Ballard D, Robinson A. Modeling embodied visual behaviors. ACM Transactions in Applied Perception 2007;4(2):11.

Swain M, Ballard D. Color indexing. International Journal of Computer Vision 1991;7:11–32.

Tatler B. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision 2007;7:1–17. [PubMed: 18217799]

Treisman A, Gelade G. A feature-integration theory of attention. Cognitive Psychology 1980;12:136.

Triesch J, Ballard D, Hayhoe M, Sullivan B. What you see is what you need. Journal of Vision 2003;3:86–94. [PubMed: 12678628]

Ullman S. Visual routines. Cognition 1984;18:97–157. [PubMed: 6543165]

Wolfe J. Guided Search 2.0: A revised model of visual search. Psychonomic Bulletin 1994;1:202–238.

Yarbus, A. Eye movements and vision. New York: Plenum Press; 1967.

Yi, W.; Ballard, D. Behavior recognition in human object interactions with a task model. IEEE international conference on Advanced Video and Signal Based Surveillance; Sydney, Australia: IEEE Computer Society; 2006.

Zihl J, von Cramon D, Mai N. Selective disturbance of movement vision after bilateral brain damage. Brain 1983;106:313–340. [PubMed: 6850272]

**Figure 1.**
Separate visual routines. When subjects have had a preview of a scene they can identify a search target's location from memory (A) but without the preview they use a correlation-based technique (B) that takes longer. One could attempt convert the remembered target's location to saliency coordinates, but not without addressing the more complicated question of how the brain manages different dynamic frames of reference.

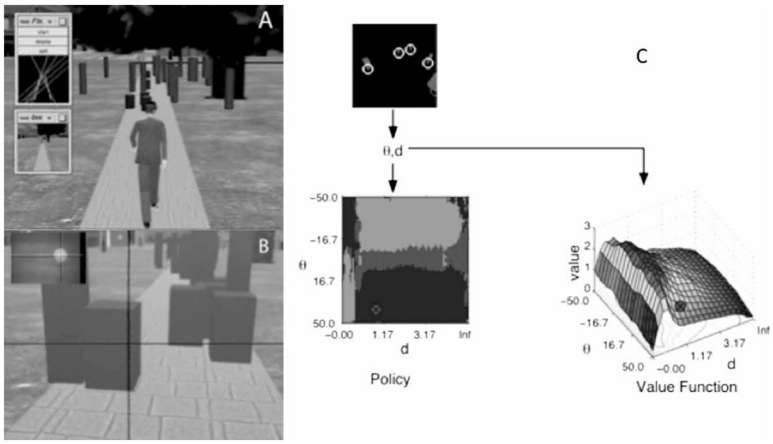


**Figure 2.**

(A) A frame from the human embedded vision system simulation showing the avatar negotiating a pavement strewn with purple litter and blue obstacles, each of which must be dealt with. The insets show the use of vision to guide the avatar through a complex environment. The upper inset shows the particular visual routine that is running at any instant. This instant shows the detection of the edges of the sidewalk that are used in navigation. The lower insert shows the visual field in a head-centred viewing frame. (B) By wearing a Head Mounted Display (HMD), humans can walk in the same environment as the avatar. (C) A basic visually guided behaviour showing steps in the use of the learnt litter cleanup Q-table. The input is a processed colour image with a filled circle on the extreme right-hand side indicating the nearest litter object as a heading angle θ and distance *d*. This state information, indicated by the circular symbol in the policy table on the lower left, is used to retrieve the appropriate action from the Q-table's policy immediately below. Light regions: Turn = −45°; grey regions: Turn = 0°; and dark regions: Turn = −45°. In this case the selected action is turn = −45°. The assumption is that neural circuitry translates this abstract heading into complex walking movements. This is true for the human avatar that has a "walk" command that takes a heading parameter. State information can also be used to retrieve the expected return associated with the optimal action, its learned Q-value, as illustrated on the lower right.

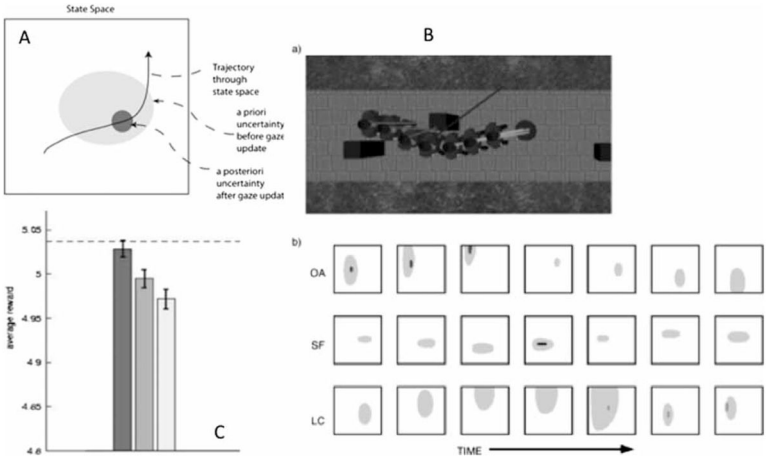


**Figure 3.**

Behaviours compete for gaze in order to update their measurements. (A) A caricature of the basic method. The trajectory through the avatar's state space is estimated using a Kalman filter that allows estimates to propagate in the absence of measurements and build up uncertainty (light grey area). If the behaviour succeeds in obtaining a fixation, uncertainty is reduced (dark grey region). The reinforcement learning model allows the value of reducing uncertainty to be calculated. (B) The top panel shows seven time steps in walking and the associated uncertainties for the state vector grey for obstacle avoidance (OA), sidewalk finding (SF), and litter pickup (LC). The corresponding boxes below show the state spaces where the a priori uncertainty is indicated in light grey and the a posteriori uncertainty is indicated in the darker grey. Uncertainty grows because the internal model has noise that adds to uncertainty in the absence of measurements. Making a measurement with a visual routine that uses gaze reduces the uncertainty. For example, for litter collection (LC), Panel 5 shows a large amount of uncertainty has built up that is greatly reduced by a visual measurement. Overall, obstacle avoidance wins the first three competitions, then sidewalk-finding, and then litter collection wins the last three. (C) Tests of the Sprague algorithm (dark) against the robotics standard round robin algorithm (light) and random gaze allocation (white) show a significant advantage over both.

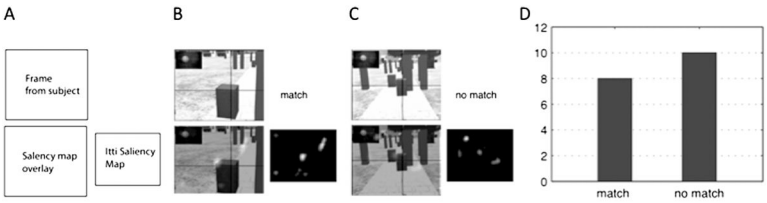


**Figure 4.**

Comparing human gaze locations to those found by the Itti saliency detector. (A) Key. The small inserts show the saliency maps that are overlaid as transparencies on the lower versions of the images. (B) Match example. (C) No match example. (D) In a sample of 18 frames, more than half show fixation locations that are not detected by the maps. The saliency program was provided by Dr. Laurent Itti at the University of Southern California. In this case, for a representative sample, only 8 out of 18 frames were labelled as matches.

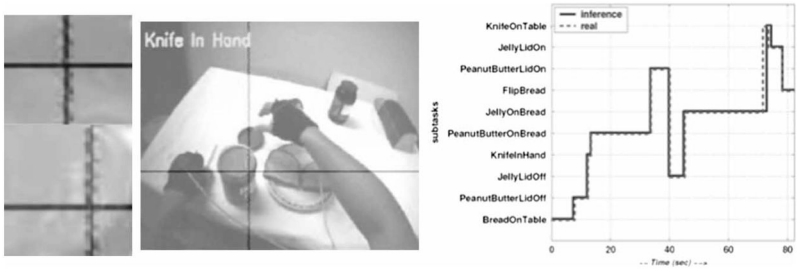


**Figure 5.**

Using the DBN to recognize steps in sandwich making. (A) Two fixations from different points in the task—(top) bread with peanut butter (bottom) peanut butter jar—appear very similar, but do not confuse the Dynamic Bayes Network (DBN), which uses task information. (B) A frame in the video of a human subject in the process of making a sandwich showing that the DBN has correctly identified the subtask as “knife-in-hand”. (C) A trace of the entire sandwich-making process showing perfect subtask recognition by the DBN.

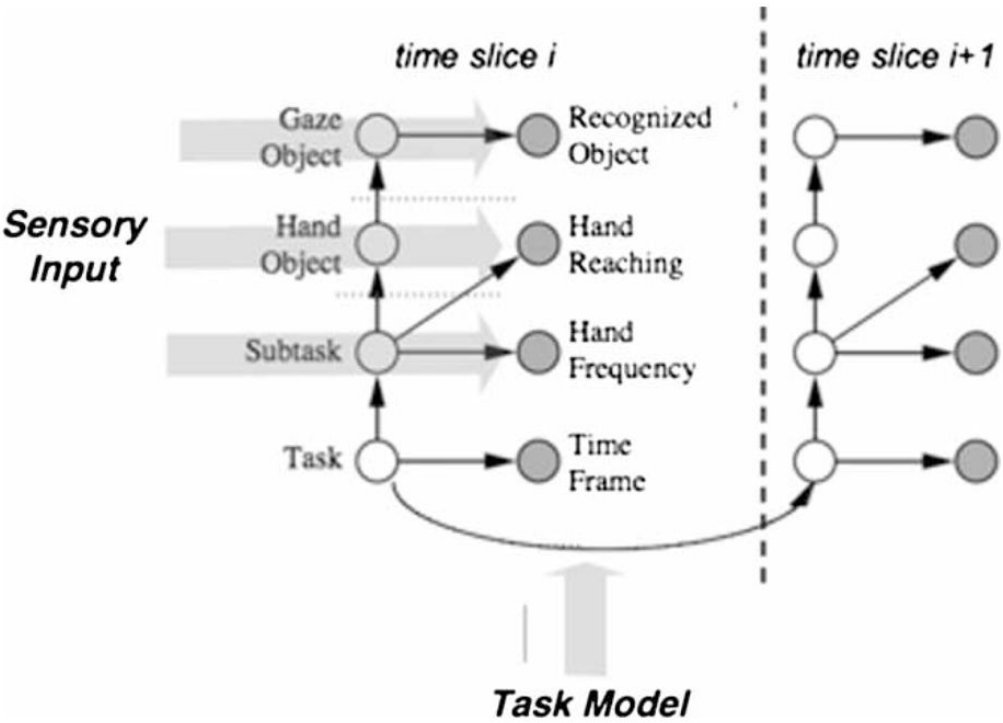


**Figure 6.**

The basic structure of the Dynamic Bayes Net (DBN) used to model sandwich making. Two time slices from the sandwich-making DBN. Visual and hand measurements provide input to the shaded nodes, the set of which at any time t comprise the measurement vector $O^t$. The rest of the nodes comprise the set $S^t$ whose probabilities must be estimated. The sequencing probabilities between subtasks are provided from a task model that in turn is based on human subject data.

**TABLE 1**

Number of states for hidden nodes in the task model (Figure 6)

| Node name | Number of states |
|---|---|
| Task | 80 |
| Subtask | 10 |
| Hand object | 4 |
| Gaze object | 5 |

## TABLE 2

Number of states for observed data nodes in the task model (Figure 6)

| Node name | Number of states |
|---|---|
| Time frame | 20 |
| Hand frequency | 2 |
| Hand reaching | 2 |
| Recognized object | 5 |